Parsimony analysis of unaligned sequence data: an exchange. Version 2 (May 2016).

Research · October 2015

2 authors:

Jan De Laet
Göteborgs Botaniska Trädgård
33 PUBLICATIONS **396** CITATIONS
SEE PROFILE

Santiago Castroviejo-Fisher
Pontifícia Universidade Católica do Rio Gran…
60 PUBLICATIONS **624** CITATIONS
SEE PROFILE

Parsimony analysis of unaligned sequence data: an exchange

10 May 2016, version 2

Introduction

Version 1

A couple of weeks ago, browsing in www.researchgate.net, I came across this paper:

Phylogenetic systematics of egg-brooding frogs (Anura: Hemiphractidae) and the evolution of direct development

SANTIAGO CASTROVIEJO-FISHER · JOSÉ M. JR. PADIAL · IGNACIO DE LA RIVA · JOSÉ P. POMBAL JR · HELIO R. DA SILVA · FERNANDO J. M. ROJAS-RUNJAIC · ESTEBAN MEDINA-MÉNDEZ · DARREL R. FROST

Zootaxa 08/2015; 4004(1):1-75. DOI:10.11646/zootaxa.4004.1.1

This paragraph on p. 8 struck me:

"***Optimality criterion and nucleotide homology***. *We chose the criterion of parsimony (unweighted) so that our phylogenetic inferences minimize ad hoc assumptions and maximize falsifiability and explanatory power of evidence (Wiley 1975; Farris 1983; Farris et al. 2001; Kluge 2001a, b, 2009; Kluge & Grant 2006; Grant & Kluge 2009). We applied parsimony to tree-alignment (Sankoff 1975; Sankoff & Rousseau 1975; Sankoff et al. 1976; Wheeler 1996) to infer the minimum number of transformation events needed to explain observed differences (including indels) in DNA sequences (Grant & Kluge 2004; 2009; Kluge & Grant 2006; Wheeler et al. 2006; Grant & Kluge 2009; Padial et al. 2014).*"

Researchgate has the nice feature that it allows direct feedback. So I asked Santiago and his coauthors a question:

And an exchange followed.

Unfortunately, graphics cannot be included in this kind of feedback, and Researchgate only provides a single font, a font that happens to be proportional. Two serious drawbacks when writing about sequence alignments. So I think it is useful to collect the ongoing online exchange in this pdf file, a format that allows some formatting to be added for readability. Other than removing some inconsistencies in interpunction, I will indicate changes and comments beyond what appeared online by putting them in green between square brackets.

This is an ongoing effort, so this pdf is expected to grow. I will document any future addition in this introduction. In the meantime, I hope that the current version may help to clarify the issues that surround parsimony analysis in a tree alignment context. Feel free to participate in the online discussion, or start a new one here.

Big thanks to Santiago for taking this up!

Jan De Laet
Veltem-Beisem
27 October 2015

Changes and additions in version 2 (10/05/2016)

Addition of Santiago's comments from 6 May 2016. Thanks to Santiago for providing me with a slightly edited version compared to the online comment.

Contribution 1: Jan - Tuesday 06/10/2015

Hi Santiago and coauthors,

I have a question.

On page 8 you say that you applied parsimony to tree-alignment to infer the minimum number of transformation events needed to explain observed differences (including indels) in DNA sequences.

But if you use minimization of transformation events as optimality criterion in a tree-alignment analysis, then you end up with a methodological breakdown. This is so because any observed sequence can then be explained by postulating just one big insertion event. For all except the most trivial datasets, that means that the data are optimally explained on any possible tree by postulating just as many insertion events as there are observed sequences. So there is no way to choose any tree over any other tree. (This is not new, I discussed it at length in the two papers at the bottom of this note; one is from 2005, the other just appeared in Cladistics).

I know that this is absurd from a biological point of view, but it follows from your stated goal to minimize transformation events in a tree-alignment analysis.

In your paper you clearly prefer some trees over other trees, so you must have been minimizing something else instead. Your paper does not contain the actual cost settings that you used in POY, so I am at a loss trying to figure out exactly what. Can you help me out?

Thanks in advance,
Best regards,

-- Jan

https://www.researchgate.net/publication/260812208_Parsimony_and_the_problem_of_inapplicables_in_sequence_data
http://onlinelibrary.wiley.com/doi/10.1111/cla.12098/abstract

---

Contribution 2: Santiago - Tuesday 06/10/2015 20:01

Hi Jan,

Thanks for the feedback . I have never had anyone commenting on my papers so this is kind of new to me. Anyways, it is very interesting to chat with other people interested in phylogenetics.

Are you Swedish? I did my PhD in Uppsala and have many emotional and professionals links to Sweden after living there for five years.

About your comment. Yes, I have read your papers. Nice contributions. Well, if I understand you comment correctly, I think I can answer your concern.

I am considering characters at the "atomic" level including indels and nucleotides. Thus, transformation series in our analysis are composed of single elements (i.e., a nucleotide or absence of nucleotide) and not groups of nucleotides and/or idels. In this sense, I think that we are actually minimizing transformation events. I agree with you that the way we expressed in the paper could be confusing. I also agree with you that it is a very naive approach to consider that all indels happen as insertion/deletions of single nucleotides. What I am still unsure if is there is a better way to do parsimony tree-alignment.

Again, if I understand your position correctly, your idea rests on interpreting parsimony as two taxon analysis (right?) as you suggest in De Laet & Smets (1998).

Best wishes,

Santiago

Contribution 3: Jan – Saturday 10/10/2015

Hi Santiago,

No, I'm not Swedish, I'm from Belgium. But as a postdoc I've been in Stockholm for a while, and for my phylogenetic research I'm now associated with Gothenburg Botanical Garden. So quite some links to Sweden here as well. I've also visited Brazil a couple of times, and also have good memories of those trips.

Thanks for your clarification and feedback. My reply here may suggest otherwise, but I do think that we have a lot of common ground.

On p. 8 you write: "We chose the criterion of parsimony (unweighted) so that our phylogenetic inferences minimize ad hoc assumptions and maximize falsifiability and explanatory power of evidence (Wiley 1975; Farris 1983; Farris et al. 2001; Kluge 2001a, b, 2009; Kluge & Grant 2006; Grant & Kluge 2009)."

I agree that parsimony minimizes ad hoc assumptions and maximizes explanatory power of evidence. But agreeing with that does not mean that I agree with all argumentation that is provided to that effect in the references that you provide. I'll concentrate on what Kluge and Grant (2006; KG6) and Grant and Kluge (2009; GK9) say about explanatory power.

They start from the philosophical principle of anti-superfluity, a parsimony principle. From that principle, they argue that explanatory power is maximized when historical transformation events, including indels, are minimized. They are explicit that this also applies in the context of tree alignments. But they don't discuss problems that might be posed by historical indel events that span multiple residues. They don't discuss how to deal with such events from a theoretical point of view, and they don't discuss how such events should be dealt with in practice, when analyzing empirical sequence data with a tree alignment program such as POY.

But from papers such as Kluge (2005) and Grant et al. (2006; full references are given at the end), it is clear that they both translate minimization of transformation events into cost set 111 for use in POY. As you pointed out, this is the same cost set that you have been using: it assigns a cost of one to transitions, a cost of one to transversions, and a cost of one to unit gaps. (I use the terminology that I also used in my 2005 and 2015 papers: a sequence such as 'a a a - - - a c c c t' has one gap that consists of three unit gaps).

This poses some fundamental problems. I'll use this small dataset of three sequences as an illustration:

```
A     a a a a c c c t
B     a a a a c c c t
C     a a a c c c a c g c t
```

With cost set 111, the optimization on the single unrooted tree for three sequences has this implied alignment:

```
A     a a a - - - a c c c t
B     a a a - - - a c c c t
C     a a a c c c a c g c t
```

It comes at a total cost of four: one base substitution and three unit gaps. So the single gap of length three is explained by postulating three distinct historical indel events, each one involving just a single position. This exposes the hidden assumption in KG6 and GK9's rationale: historical indel events never involve more than one nucleotide at a time. I agree that this is a naive assumption. But it sits deeply embedded in the core of KG6 and GK9's view of parsimony analysis: in their view, each base substitution and each unit gap stands for a distinct historical event.

There is still a deeper problem. KG6 and GK9 rely on the philosophical notion of anti-superfluity. But properly applied in this context, that principle implies the following: when there is a choice between an explanation that involves three transformation events and an explanation that only involves a single transformation event, then the explanation with only a single transformation event should be preferred. In other words, If I can explain a single gap of length three with only one transformation event, I should not postulate three such events. And this should also apply during analysis. So KG6 and GK9's position not only involves an unrealistic biological assumption, the theoretical core of their position itself is self-contradictory in its application of anti-superfluity, the central concept on which that core is built (I've discussed this in my recent paper in Cladistics).

You say that you are considering 'characters at the "atomic" level including indels and nucleotides. Thus, transformation series in our analysis are composed of single elements (i.e., a nucleotide or absence of nucleotide) and not groups of nucleotides and/or idels. In this sense, I think that we are actually minimizing transformation events'.

I agree with that, but only as far as it goes: you are minimizing some abstract concept of transformations (and doing so postulating some abstract and non-standard notion of transformation series). But such transformations have no sensible biological meaning, and they are not the kind of unique historical events that KG6 and GK9 rely on in their theoretical framework (even if they give up that meaning when it comes to empirical work or recommendations). So I don't think it makes sense to refer to KG6 and GK9 as a rationale for 111.

So, are there alternatives?

What happens, for example, if, within KG6 and GK9's view of parsimony, you give up the assumption that indels only affect single nucleotides at a time? That leads to giving up 111. The result is the methodological breakdown that I mentioned in my previous post.

One might take a more pragmatic stance: first use 111 to get trees, then use those trees to infer where indels might have affected multiple positions (to be sure, I've never seen this position defended in papers, I'm just exploring possibilities). This might give decent results in practice, but as a method it is inconsistent: during the analysis it is assumed that indels affect only single nucleotides at a time, after the analysis, this assumption is given up. It may be useful to discuss an example. Consider this dataset:

```
A     a a a
B     a a a a a a
C     a a a a a a a a a
D     a a a a a a a a a a a a
```

Analysis with cost set 111 leads to the following total costs on the three different unrooted trees for four terminals:

```
(A B)(C D)) -> total cost 9
(A C)(B D)) -> total cost 15
(A D)(B C)) -> total cost 15
```

So, during analysis, ((A B)(C D)) is preferred as the best explanation because it minimizes historical events under the assumption that historical indel events only affect single nucleotides. After the analysis, it is then concluded that the best explanation actually involves only three historical indel events (all three involving a subsequence of length three). But, admitting that indel events can involve multiple nucleotides, the two other trees can also explain the data with only three historical indel events (but involving subsequences of

different lengths). The net result is that equally good explanations under a realistic assumption are not considered because they have previously been rejected during an analysis that was performed under an unrealistic assumption.

One might still be more pragmatic: 'I don't care what the underlying rationale is, operationally 111 has again and again proven to give decent results, so I'll stick to it'. Or '111 is simple, that's sufficient'. That would come close to arguments of simplicity that Wheeler at times has voiced.

In 2003 and 2005 I've argued from a completely different perspective, one that you indeed could say was inspired by my 1997 view of parsimony as two item-analyis (I only explicitly called it such in 1998, but the ideas are there). It leads to a view that using 111 in POY amounts to a specific kind of differential weighting of evidence (I've elaborated that in my recent paper in Cladistics). To obtain (an approximation of) equally weighted evidence, one should use cost set 3221 (gap opening cost 3, transition and transversion cost 2, gap extension cost 1). I'll try to expand a bit on that later this weekend.

Best

Jan

De Laet, J., and Smets, E. 1998. On the three-taxon approach to parsimony analysis. Cladistics 14: 363-381.

De Laet, J. 1997. A reconsideration of three-item analysis, the use of implied weights in cladistics, and a practical application in Gentianaceae. Dissertation. Available at wwww.anagallis.be or in ResearchGate.

De Laet, J. 2003. When one and one is not two: parsimony analysis of sequence data. XXIIth Meeting of the Willi Hennig Society. New York Botanical Garden, 20 July – 24 July. Abstract appeared in Cladistics 20: 81 (2004). Also available at www.anagallis.be.

Grant, T., Frost, D.R., Caldwell, J.P., Gagliardo, R., Haddad, C.B., Kok, P.J.R., Means, D.B., Noonan, B.P., Schargel,, W.E., Wheeler, W.C., 2006. Phylogenetic systematics of dart-poison frogs and their relatives (Amphibia; Athesphatanura: Dendrobatidae). Bull. Am. Mus. Nat. Hist. 299.

Kluge, A.G., 2005. What is the rationale for 'Ockham's Razor' (a.k.a. parsimony) in phylogenetic inference?. In: Albert, V. (Ed.), Parsimony, Phylogeny, and Genomics. Oxford University Press, Oxford, pp. 15–42.

Contribution 4: Santiago - Wed 14/10/2015

Hi Jan,

Thanks for taking your time to explain your position. As you said, instrumentalist justifications such as ‘I don’t care what the underlying rationale is, operationally 111 has again and again proven to give decent results, so I’ll stick to it’. Or ‘111 is simple, that’s sufficient’.
are empty and should not be used. I would greatly appreciate if you find the time to elaborate a bit more on the 3221 cost regime.

In any case, I want to study your papers in detail to fully grasp your proposal. It sounds very interesting. I will be in contact with you.

Cheers,

santiago

Contribution 5: Jan - Sat 24/10/2015

Hi Santiago,

You wondered if there's a better way to do parsimony analysis in a tree alignment context than to assume that indel events affect only single nucleotides at a time, or than to stick to purely instrumentalist justifications. I think there is, and it indeed involves cost set 3221 for use in POY. Before getting to that cost set, a lengthy introduction may useful though.

To illustrate some concepts and ideas, I will mainly be using this hypothetical dataset of four observed sequences (that they have been put in a dataset for phylogenetic analysis means that they are hypothesized to be orthologous):

```
Dataset D1
A     gggaaaacccggg
B     gggaaaaaatttggg
C     gggaaaaaaaaaacccggg
D     gggaaaaaaaaaaaatttggg
```

TREE ALIGNMNENTS

A tree alignment is a concept that is due to David Sankoff (see Sankoff 1975, Sankoff and Cedergren 1983). For a dataset of unaligned sequences, such as D1, a tree alignment consists of (1) a tree with the observed sequences at the tips and reconstructed sequences at the inner nodes, and (2) a multiple alignment of observed and reconstructed sequences alike. (The multiple alignment of the observed sequences that is obtained by deleting the inner nodes from a tree alignment is called an implied alignment). A nice representation of a tree alignment is a drawing of that tree in which each node is labelled with the corresponding row of that multiple alignment. Unfortunately I can't do this here because ResearchGate doesn't permit a non-proportional font or the use of graphics here. So I'll have to use a somewhat less clear representation, with trees in parenthetical notation and numbered inner nodes [I'll stick to it in this formatted version]. For four terminals there are three different unrooted trees:

```
T1: (1:(A B) 2:(C D))
T2: (3:(A C) 4:(B D))
T3: (5:(A D) 6:(B C))
```

Consider unrooted tree T1, (1:(A B) 2:(C D)). The single inner branch determines partition AB|CD. The inner node at the AB-side is labelled node 1, the inner node at the CD side is called node 2. In the same way, the inner nodes of trees T2 and T3 are called 3, 4, 5 and 6.

Given this convention, the following multiple alignment fully determines a tree alignment for dataset D1 on tree T1:

```
Tree alignment D1T1A1 (first tree alignment for dataset D1 on tree T1)
A     gggaaaa--------cccggg
B     gggaaaaaa------tttggg
C     gggaaaaaaaaaa--cccggg
D     gggaaaaaaaaaaaatttggg
1     gggaaaaaa------tttggg
2     gggaaaaaaaaaa--tttggg
```

It looks a bit garbled in the single font that ResearchGate provides, but it should be clear enough. If not, it might be useful to paste it in some program that allows a non-proportional font (Courier, for example). [A superfluous paragraph in this formatted version].

The length of tree alignment D1T1A1 is 21, so there are 21 positional characters. In total, these have six substitutions: two in position 16, two in position 17, and two in position 18. Three of those are in the terminal branch leading to A, the other three in the terminal branch leading to C. The length differences of the observed sequences are explained by three indel events: an indel event of a subsequence of length two along the branch leading to A, an indel of length four along the inner branch, and an indel of length two along the branch leading to D. Whether such an indel is an insertion or a deletion depends on how the tree would be rooted. So the explanation of the observed sequences that is provided by D1T1A1 requires a total of nine evolutionary events or transformations.

Here's another tree alignment for D1 on T1:

```
Tree alignment D1T1A2
A      gggaaaa--------cccggg
B      gggaaaaaa------tttggg
C      gggaaaaaaaaaa—cccggg
D      gggaaaaaaaaaaaatttggg
1      gggaaaaaa------cccggg
2      gggaaaaaaaaaa—cccggg
```

D1T1A1 and D1T1A2 only differ in the reconstructed sequences at positions 16, 17, and 18: both inner nodes have a 't' there in D1T1A1, but a 'c' in D1T1A2. And just as in D1T1A1, six subsitutions are required in these positions, and so the total number of evolutionary events in both tree alignments is the same. The difference is that, in D1T1A2 B, the subsitutions occur in the terminal branches that lead to B and D, not in the branches that lead to A and C.

Here's a tree alignment for D1 on T2, also with 21 positions:

```
Tree alignment D1T2A1
A      gggaaaa--------cccggg
B      gggaaaaaa------tttggg
C      gggaaaaaaaaaa--cccggg
D      gggaaaaaaaaaaaatttggg
3      gggaaaaaaaaaa--cccggg
4      gggaaaaaaaaaa--tttggg
```

This tree alignment requires only three substitutions, all three along the single inner branch: one in position 16, one in position 17, and one in position 18. The length differences still require only three indel events, but of different lengths compared to D1T1A1 and D1T1A2: an indel of length six along the branch to A, an indel of length four along the branch to B, and an indel of length two along the branch to A. So in total, this tree alignment only requires six evolutionary events: three substitutions and three indel events.

One could be tempted to prefer D1T2A1 - and hence tree T2 – over D1T1A1 and D1T1A2 because it requires less evolutionary events. But using that criterion, the three following tree alignments perform even better:

Tree alignment D1T1A3

```
A ------------------------------------------------------------gggaaaacccggg
B ---------------------------------------gggaaaaaatttggg-------------
C --------------------gggaaaaaaaaaacccggg--------------------------
D gggaaaaaaaaaaatttggg-----------------------------------------------
1 ------------------------------------------------------------------
2 ------------------------------------------------------------------
```

Tree alignment D1T2A2

```
A ------------------------------------------------------------gggaaaacccggg
B ---------------------------------------gggaaaaaatttggg-------------
C --------------------gggaaaaaaaaaacccggg--------------------------
D gggaaaaaaaaaaatttggg-----------------------------------------------
3 ------------------------------------------------------------------
4 ------------------------------------------------------------------
```

Tree alignment D1T3A1

```
A ------------------------------------------------------------gggaaaacccggg
B ---------------------------------------gggaaaaaatttggg-------------
C --------------------gggaaaaaaaaaacccggg--------------------------
D gggaaaaaaaaaaatttggg-----------------------------------------------
5 ------------------------------------------------------------------
6 ------------------------------------------------------------------
```

These three tree alignments can 'explain' the observations by postulating only four indel events (of lengths 21, 19, 15, and 13), an explanation that is optimal on every possible tree. They illustrate the methodological breakdown that follows when minimizing equally weighted evolutionary events in a tree alignment context and under the assumption that single indel events can involve more than one nucleotide. But such tree alignments actually explain nothing at all. Given the prior hypothesis that the observed sequences are orthologous, tree alignment D1T1A1 can, for example, explain the shared presence in B and D of a stretch of three t's near the end of their sequence (they inherited it from their common ancestor). Trivial tree alignments D1T1A3, D1T2A2, D1T3A1 cannot explain such shared presences.

I have argued (De Laet 2005, 2015) that a proper generalization of parsimony to tree alignments should focus directly on what alternative hypotheses (tree alignments) can explain about observed empirical data, not on evolutionary events that are required to that effect.

PARSIMONY AND EXPLANATION

When it comes to explanation, the basic observation is that "genealogies provide only a single kind of explanation. A genealogy does not explain by itself why one group acquires a new feature while its sister group retains the ancestral trait. … A genealogy is able to explain observed points of similarity among organisms just when it can account for them as identical by virtue of inheritance from a common ancestor." (Farris 1983, p. 13, as quoted by Farris 2006, pp. 825-826).

As an illustration, consider hypothetical dataset D2, a dataset with just a single morphological character:

Dataset D2

```
A  0
B  0
C  1
D  1
```

On tree T1, the shared presence of state zero in A and B (one observed point of similarity) and the shared presence of state 1 in C and D (another observed point of similarity) can simultaneously be explained as due to common descent. This explanation requires a character state reconstruction that assigns state 0 to inner node 1, and state 1 to inner node 2.

This kind of explanation is is independent of the position of the root. Assume for example that the tree is rooted along the branch that leads to A. In that case, state 0 is plesiomorphic and A and B inherited it from the common ancestor of all four terminals. Apomorphic state 1 arose along the branch that leads to the common ancestor of C and D, and C and D inherited it from that common ancestor. Alternative rootings will differ in assessment of plesiomorphy and apomorphy, but in all possible scenarios both the observed point of similarity in state zero and the observed point of similarity in state one can be explained by common ancestry.

On tree T2, the shared presence of state 0 in A and B and the shared presence of state 1 in C and D cannot simultaneously be explained by common ancestry. It is possible to explain the shared presence of state 0 in A and B in that way (with a reconstruction that assigns state 0 to both inner nodes), but then the shared presence of state 1 cannot be so explained. It is then an instance of homoplasy or unexplained shared similarity. Alternatively, it is possible to explain the shared presence of state 1 in C and D by common ancestry (with a reconstruction that assigns state 1 to both inner nodes), but then the shared presence of state 0 in A and B can no longer be so explained. The same is true for tree T3: either the shared presence of state 0 in A and B or the shared presence of state 1 in C and D can be explained by common ancestry. But not both.

For the simple character of dataset 2, there is at most a single unexplained point of shared similarity (none on T1, one on trees T2 and T3). With more terminals, there can be more such instances within a single character, and these may be logically interdependent. When minimizing homoplasy or unexplained similarity - as a means to maximize explained observed similarity - not all instances of pairwise homoplasy should then be counted, but only instances that are logically independent.

As Farris (2006, p. 826) put it (mainly by citing from his 1983 paper): ‘It is common for homoplasies to be logically interdependent (Farris, 1983, p. 20): “Suppose that a putative genealogy distributes [the 20 terminals showing feature X] into two distantly related groups A and B of ten terminals each. There are 100 distinct two-taxon comparisons of members of A with members of B, and each of those similarities in X considered in isolation comprises a homoplasy… [But if] X is identical by descent in any two members of A, and also in any two members of B, then the A-B similarities are all homoplasies if any one of them is.” But fortunately it is easy to count mutually independent homoplasies (Farris, 1983, p. 20): “If a genealogy is consistent with a single origin of a feature, then it can explain all similarities in that feature as identical by descent. A point of similarity in a feature is then required to be a homoplasy only when the feature is required to originate more than once on the genealogy. A hypothesis of homoplasy logically independent of others is thus required precisely when a genealogy requires an additional origin of a feature. The number of logically independent ad hoc hypotheses of homoplasy in a feature required by a genealogy is then just one less than the number of times the feature is required to originate independently.’

The same is true when directly counting explained points of similarity, in order to maximize them directly (De Laet 1997, pp. 66-67). Using Farris' example, if terminals A1, A2, and A3 are all three members of group A, then there are three observed points of similarity among those three terminals that can be explained by inheritance and common descent: the similarity between A1 and A2, the similarity between A1 and A3, and the similarity between A1 and A3. But these are logically interdependent: if any of two out of those three hold, then the third follows by necessity. When such logical interdependencies are properly taken into account, the number of explained shared similarities in such a character on a tree varies directly with the number of steps that are required for that character on that tree: every additional step amounts to one less independent observed point of similarity that can be explained.

As each independent explained point of similarity involves two terminals, parsimony analysis can be characterized as two-item analysis (see De Laet 1997, p. 66-67): it identifies the trees that maximize the number of independent observed pairwise points of similarity that can simultaneously be explained by inheritance and common descent. Observed pairwise similarities are the atomic units of empirical comparative content of a dataset, and the units with which to measure the explanatory power of a tree with optimized characters.

COMPOSITIONAL SEQUENCE HOMOLOGY

Rather than to rely on the relationship with number of steps, independent explained similarities in a character on a tree can be counted directly.

Doing this for a general (unordered) morphological character, that number is equal to the number of observations in the character minus one minus the number of steps that the character has on the tree (this is a straightforward generalization of the binary case, discussed in De Laet 1997, pp. 66-67). For the character of dataset D2, there are four observations (one in each terminal; missing information would not be counted as an observation). On a tree that has an AB|CD partition, the character has one step and the above number equals two (4 – 1 - 1). So there are two independent explained shared observed similarities. The first one is the single similarity in state 0, the second the single observed similarity in state 1. On a tree that does not have this partition, two steps are required and the above number amounts to one: either the similarity in state 0 or the similarity in state 1. The tree with the AB|CD partition is preferred because it can explain one more independent observed point of similarity.

When applying this point of view to a set of unaligned sequences, things get more complicated because there are no predefined positions and positional characters to which the above calculation can be applied. But for any given tree alignment for that set of unaligned sequences, it is possible to count how many independent points of sequence similarity in base composition can be explained by common descent and inheritance.

That number is equal to the total number of nucleotides in the observed sequences minus the total number of subcharacters in the tree alignment minus the total number of base substitutions within those subcharacters (De Laet 2005, pp. 107-108; see also De Laet 2015, p. 552). I have called this the compositional component of sequence homology (De Laet 2005, p. 106; see also De Laet 215, p. 551).

In the above expression, a subcharacter of a tree alignment is a region in the tree where a particular position is applicable. Position 13 of D1T1A1, for example, has two nucleotides: an 'a' in terminal C, and an 'a' in terminal D. The inner node that connects terminals C and D (inner node 2) also has a nucleotide in that position. Therefore the two observed nucleotides at that position are in the same subcharacter in this tree alignment. Within this subcharacter, there are no substitutions. As another example in that same tree

alignment, there are four nucleotides in position 18: a ‘c’ in A and C, and a ‘t’ in B and D. The two inner nodes that connect these terminals also have a nucleotide at that position. So there is a single subcharacter at position 19 as well. In this one, there are two substitutions.

For an example where a single position has more than one subcharacter, consider position 10 of the following tree alignment of dataset D1 on tree T1:

```
Tree alignment D1T1A4
A     gggaaaa--------cccggg
B     gggaaaaaa------tttggg
C     gggaaaaaaaaaa--cccggg
D     gggaaaaaaaaaaaatttggg
1     gggaaaaaa------tttggg
2     gggaaaaaa------tttggg
```

That position has two nucleotides: an ‘a’ in terminal C and another ‘a’ in terminal D. But in this tree alignment (a suboptimal one, to be sure), the inner node that connects these observed nucleotides (inner node 2) does not have a nucleotide at that position. Therefore, these two observed nucleotides are not directly comparable. They are in two different subcharacters or regions of applicability. Trivial regions, for that matter: each subcharacter in position 10 of this tree alignment has only a single nucleotide.

Within any given subcharacter of a tree alignment, the number of independent explained similarities in base composition can be obtained by applying the formula for a regular unordered character (number of observations minus one minus steps), but restricted to the terminals that participate in the subcharacter: the number of nucleotides within the subcharacter minus one minus the number of transformations within the subcharacter (see below for some examples). When this is summed over all subcharacters of a tree alignment, the above grand total for the complete tree alignment is obtained: the total number of nucleotides in the observed sequences minus the total number of subcharacters in the tree alignment minus the number of substitutions within subcharacters.

Dataset D1, for example, has 68 observed nucleotides (the sum of the lengths of the observed sequences). Tree alignment D1T1A1 has 21 positions, and in each of these positions the observed nucleotides are in a single region of applicability. So D1T1A1 has 21 subcharacters. Within the subcharacters, there are six substitutions: two in the single subcharacter at position 16, two in the single subcharacter at position 17, and two in the single subcharacter at position 18. So the total measure of compositional homology equals 68 – 21 – 6 = 41.

Most of these 41 explained points of similarity in base composition are in the stretches of nucleotide ‘a’ and the stretches of nucleotide ‘g’ in the observed sequences. Take for example the first position: a single subcharacter that comprises all four terminals. In this subcharacter, all terminals in it have an observed ‘g’. This amounts to three independent instances of compositional homology (if, for example, the ‘g’ in A and B is homologous, the ‘g’ in A and C is homologous, and the ‘g’ in C and D is homologous, then by necessity all other pairwise homologies follow). This number (3) can be obtained as the number of observed nucleotides in the subcharacter minus one minus the number of substitutions or steps within the subcharacter: 4 – 1 – 0 = 3.

Or position 13 of D1T1A1, with a single subcharacter that comprises just C and D. Both terminals have an observed ‘a’ there, which amounts to a single instance of compositional homology. Using the above formula, this is obtained as 2 observed nucleotides minus 1 minus 0 substitutions.

Summing over all subcharacters in which no substitutions occur (all subcharacters except those at positions 16, 17 and 18), 38 such independent instances of compositional homology can be counted.

The other three are in the subcharacters at positions 16, 17, and 18, the only subcharacters with substitutions in this tree alignment. Given that both inner nodes have a reconstructed 't' for each of these three subcharacters, the observed shared similarity in each of the three nucleotides 't' near the end of the observed sequences of B and D is homologous: their presence can be explained by common ancestry. The similarity of three nucleotodes 'c' near the end of the observed sequences of A and C, on the other hand, cannot be explained by common ancestry. In total, this amounts to three independent explained points of similarity in these subcharacters: one in position 16, one position 17, and one in position 18. (Applying the formula for any of these three subcharacters: 4 observed nucleotides minus 1 minus 2 substitutions).

This result can be contrasted with tree alignment D1T2A1 for that same dataset. Just as D1T1A1, tree alignment D1T2A2 has 21 positions and exactly one subcharacter at every position. But in total, there are only three substitutions: one in the subcharacter at position 16, one in the subcharacter at postion 17, and one in the subcharacter at postion 18. This amounts to the following grand total of explained similarity in base composition: 68 – 21 – 3 = 44.

The difference with D1T1A1 is 3 (44 - 41). Given the similarity between tree alignments D1T1A1 and D1T2A1, it is easy to verify this statement: all points of similarity in base composition that DT1A1 can explain, can also be explained by D1T2A1. The difference is that D1T2A1 can explain three more similarities. These three additional explained similarities reside in the subcharacters at positions 16, 17, and 18.

SUBSEQUENCE HOMOLOGY

Sequence homology in a tree alignment is not captured completely by just looking at nucleotide level homology within positions. This is illustrated using dataset D3:

Dataset D3

```
A  aaaaaa
B  aaaaaa
C  aaagggaaa
D  aaagggaaa
```

Consider these two tree alignments:

Tree alignment D3T1A1

```
A   aaa---aaa
B   aaa---aaa
C   aaagggaaa
D   aaagggaaa
1   aaa---aaa
2   aaagggaaa
```

Tree alignment D3T2A1

```
A   aaa---aaa
B   aaa---aaa
C   aaagggaaa
D   aaagggaaa
3   aaagggaaa
4   aaagggaaa
```

In D3T1A1, the shared presence of subsequence ‘ggg’ in the middle of the observed sequences of C and D can be explained by common ancestry. As can its absence in A and B. In D3T2A1, that shared presence can still be explained by common ancestry, but its absence in A and D can no longer be so explained. The difference in explanatory power (one less explained point of similarity in D3T2A1) is exactly matched by the difference in number of indel events that both tree alignments require: one indel event, of subsequence ‘ggg’, along the inner branch of tree T1 for D3T1A1; two such events in tree T2 for D3T2A1(one along the branch leading to A, the other along the branch leading to B).

In this simple case, the subsequence involved is identical in the two terminals where it is observed. But that does not need to be the case. Consider dataset D4:

Dataset D4

```
A  aaaaaa
B  aaaaaa
C  aaagggaaa
D  aaagtgaaa
```

D3 and D4 are identical except for the middle position of the observed sequence of D: in D3 it is a ‘g’, in D4 a ‘t’.

Consider these two tree alignments for D4:

Tree alignment D4T1A1

```
A   aaa---aaa
B   aaa---aaa
C   aaagggaaa
D   aaagtgaaa
1   aaa---aaa
2   aaagggaaa
```

Tree alignment D4T2A1

```
A   aaa---aaa
B   aaa---aaa
C   aaagggaaa
D   aaagtgaaa
3   aaagggaaa
4   aaagggaaa
```

Even if the three middle positions in the sequences of C and D are no longer identical, the shared presence of a subsequence at those positions can still be explained by common ancestry. I have called this subsequence homology, a component of sequence homology that cannot be reduced to homology of observed nucleotides within orthologous subsequences or subcharacters (De Laet, 2005, p. 106; see also De Laet 2015: 551-552).

This can be illustrated by rooting tree T1 for example along the branch that leads to D (which involves the assumption that D is a proper outgroup for this set of terminals). In that case, the absence in A and B of the middle subsequence that is observed in C and D was inherited from their common ancestor, and it provides a synapomorphy for A and B. One, moreover, that cannot be expressed in terms of composition of an observed subsequence in those terminals.

In more complicated datasets and tree alignments, the subsequences that are involved in indel events along different branches may not fully coincide. There are partially overlapping indels along different

branches in such cases. But even then the number of indel events can be used to compare this kind of explained similarity between any two different tree alignments: whenever the first tree alignment has an indel event that is not present in a second one, the first tree alignment has an unexplained similarity across the branch with that indel, compared to the second one: depending on the root, either a subsequence that was present got lost, or a subsequence that was not present before was gained. The same holds the other way around. The net balance of indels between the two tree alignments is then a proper measure with which their subsequence homology can be compared.

ADDING IT UP

So compositional homology in a tree alignment is directly measured by the number of nucleotides in the observed sequences minus the number of subcharacters in the tree alignment minus the number of substitutions within subcharacters. And the number of indel events of subsequences provides a measure with which the amount of subsequence homology can be compared.

Adding it up, when comparing two tree alignments for a set of observed sequences, the tree alignment with the higher amount of total (equally weighted) sequence similarity that can be explained as homology is the one with the higher value for this aggregate number: the number of nucleotides in the observed sequences minus the number of subcharacters in the tree alignment minus the number of substitutions within subcharacters minus the number of indel events. This holds in general and does not depend on the aligned length.

The aggregate number (and hence explained shared sequence similarity) is maximal for the tree(s) and tree alignment(s) for which the sum of subcharacters, substitutions, and indel events is minimal.

POY

In POY, the substitution cost can be used to minimize substitutions. And indel events that span multiple nucleotides can in principle be minimized by using a positive gap opening cost and a zero gap extension cost. But POY does not provide a cost parameter to minimize subcharacters. So it may look like POY cannot be used to maximize sequence similarity that can be explained as homology. But there is a practical work-around.

First some background on tree alignment algorithms. Exact algorithms to find an optimal tree alignment on a given tree are so computationally complex that heuristic approximations are unavoidable in practice. This is where POY's algorithms such as DO (direct optimization) and iterative-pass optimization come in. These algorithms never consider more than three sequences at a time when reconstructing a sequence at an inner node.

Consider for example iterative-pass optimization, an approach due to Sankoff et al. (1973). Using tree T1 as an example, this is a general description. The algorithm starts out with the calculation of an initial reconstructed sequence for the two inner nodes. Next, the algorithm enters an iterative process in which it tries to improve these initial reconstructions by repeatedly revisiting the inner nodes. Assume that it first revisits inner node 1. This inner node has three incident branches: the branches leading to observed sequences A and B and the branch leading to inner node 2. Using observed sequences A and B and the reconstructed sequence at inner node 2, a new reconstructed sequence for node 1 is calculated (the so-called median sequence). Next it will consider node 2 and recalculate the reconstructed there using the sequences at nodes C, D, and the new reconstructed sequence at inner node 1. This is repeated until no more improvements can be made. That is, until the cost stabilizes.

Computational complexity is such that it is doable to calculate an exact median sequence during any revisit of an inner node. Exactness here means that that median sequence is guaranteed to be optimal. Given the input, that is: the sequences at the ends of the three incident branches. In terms of optimizing the cost, this algorithm definitely performs better than simple DO (at the expense of longer execution time). But still it never considers interactions across longer distances than subtrees that consist of one inner node and its three neighbouring nodes. As a result, exact optimality can only be guaranteed for datasets up to three terminals. Beyond that, it provides a heuristic approximation. This is a general conclusion and does not depend on the cost parameters being used.

Back to maximizing explained similarity. It can be shown that the 3221 cost regime measures, up to a constant, (twice) the number of subcharacters, substitutions, and indel events for (sub)trees that consist of one inner node and three neighbouring nodes (De Laet 2005, p. 109; see also de Laet 2015, p. 562-563). This means that explained sequence similarity is guaranteed to be optimal when using that cost set on such (sub)trees. For larger trees or subtrees, that guarantee of optimality no longer holds. But POY's algorithms don't guarantee optimality for such larger trees or subtrees to start with. So, in practice, 3221 is the best possible heuristic approximation for maximization of explanatory power using the algorithms that are available in POY.

A discussion of an analysis of dataset D1 using POY might be useful at this point, but I reckon that this post is already way too long. So I'll keep that for some other time. In the meantime I hope that this post may be useful as a clarification of my position on these issues.

Best

-- Jan

REFERENCES

De Laet, J., 1997. A reconsideration of three-item analysis, the use of implied weights in cladistics, and a practical application in Gentianaceae. Dissertation. Available at www.anagallis.be or in ResearchGate.

De Laet, J. 2005. Parsimony and the problem of inapplicables in sequence data. Pp. 81-116 in Albert, V. A. (ed.), Parsimony, phylogeny and genomics. Oxford University Press.

De Laet, J., 2015, Parsimony analysis of unaligned sequence data: maximization of homology and minimization of homoplasy, not minimization of operationally defined total cost or minimization of equally weighted transformations. Cladistics, 31: 550–567. doi: 10.1111/cla.12098.

Farris, J. S., 1983. The logical basis of phylogenetic analysis. Pp. 7-36 in: Platnick, N. I., Funk, V. A. (Eds.), Advances in Cladistics II. Columbia University Press, New York.

Farris, J. S., 2008. Parsimony and explanatory power. Cladistics, 24: 825–847.

Sankoff, D. 1975. Minimal mutation trees of sequences. SIAM J. Appl. Math. 28: 3542.

Sankoff, D., and Cedergren, R. J. 1983. Simultaneous comparison of three or more sequences related by a tree. Pp. 253-263 in Sankoff, D., and Kruskal, J. (eds.), Time warps, string edits, and macromolecules. The theory and practice of sequence comparison.
CSLI Publications, Stanford, California (1999 reprint).

Sankoff, D., Morel, C., and Cedergren, R. J. 1973. Evolution of 5S RNA and the nonrandomness of base replacement. Nature (New Biology) 245: 232-234.

Contribution 6: Santiago – Friday 06/05/2016

Dear Jan,

I am sorry it has taken me so long to answer. Thanks to your detailed correspondence, I was able to better understand your ideas. Your papers in *Cladistics* are quite technical and I have to admit that I was having difficulties to follow some of your arguments. I am also what you can call a "slow thinker" and like to have my time to reflect on things. In February, I was doing field work in the mountain rainforests of Peru. For almost a month, I was hiking every day looking for amphibians and squamates, living in a tent, and cooking by the camp fire. It was a very nice setting and scenario to think calmly about your ideas. In any case, here are my two cents on the topic.

In a nutshell, your concern arises from the fact that the assumption "indels never involve more than one nucleotide at a time" is (i) an unrealistic biological assumption and (ii) logically irreconcilable with the anti-superfluity principle used by Kluge and Grant.

Below, I will expand on these two points and how I think they should be considered in the light of character independence.

An unrealistic biological assumption

I guess that nobody with some basic notions on biology will deny that "indels never involve more than one nucleotide at a time" is false. However, you do not mention that this false assumption actually implies other equally important, with regards to their implications to phylogenetics, statements such as:

- Indels do not always involve more than one nucleotide at a time
- Some indels, even those affecting more than one nucleotide, are the result of transformations of one nucleotide at a time
- Some indels affecting more than one nucleotide are the result of a combination of transformations of one nucleotide at a time and several nucleotides at a time

Following the same logic, we can also consider false the assumptions "mutations never involve more than one nucleotide at a time" and "transformations of morphological characters never involve more than one character at a time", and the same sort of statements outlined above also follows from these false assumptions.

You criticized the arguments of Kluge and Grant for considering "indels never involve more than one nucleotide at a time" but at the same time the same critic can be applied to your proposal because it assumes that "indels always involve more than one nucleotide at a time", which arguably is also biologically unrealistic. I see both statements as extremes of a continuum, because, as explained above, there are all sorts of situations in between. In other words, the number of possible permutations between single character events (*e.g.*, indels, mutations, transformation of phenotypic character) and single events affecting multiple characters (*e.g.*, indels affecting more than one nucleotide at a time, mutations affecting more than one nucleotide at a time, transformation of more than one phenotypic character at a time) within a real dataset is absurdly large! I guess that the only biologically plausible assumption in this context is what my host at AMNH, Darrel Frost, used to say about evolution "shit happens", meaning that not only

many of those possibilities are evolutionary plausible but that they may have happened, at least once, during the ~3500 million years on life on Earth.

Logically irreconcilable with the anti-superfluity principle

You said that if properly applied, the anti-superfluity principle used by Kluge and Grant to justify their view favors your position that "indels always involve more than one nucleotide at a time" because the explanation with less transformation events should always be favored. OK, I see the beauty of that and also that applying the same antisuperfluity principle *naively* takes us to the trivial solution that every difference (not only indels) is most parsimoniously explained by a single transformation event affecting all the characters. This would be similar, although a more general case because it would also account for mutations, to the example you provide in page 9 of our exchange.

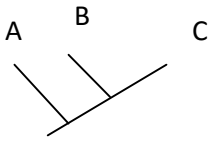
What I call a *naïve* application of the anti-superfluity principle is the idea of applying this principle in a vacuum, without considering anything else. Phylogenetic analyses are performed within a context of auxiliary assumptions or principles, for example we consider that evolution is hierarchical and that characters are independent. The second consideration is most crucial to the anti-superfluity principle. When not considered, it leads to trivial solutions. I never read any contribution by Grant and/or Kluge defending this *naïve* use of the anti-superfluity principle. To the contrary, Grant & Kluge (2004) *Cladistics*, 20, 23-31 devote page 26 to the importance of character independence.

To avoid this type of trivial solutions or *naïve* use of the anti-superfluity principle (*i.e.*, where a single transformation is invoked to explain all differences between characters) you suggest embracing explanation of similarities computed as: (number of nucleotides in the observed sequences) – (number of subcharacters in the tree alignment) – (number of substitutions within subcharacters) – (number of indel events). However, you count indel events applying the assumption "indels always involve more than one nucleotide at a time" by defining subsequence homology so, if I understand it correctly, your approach does not circumvent the *naïve* use of the anti-superfluity principle at least when applied to indels expanding more than one nucleotide. This also leaves the unanswered question, why one should not count mutations the same way you are counting indels? This is, for a subsequence of nucleotides, the number of nucleotides that are involved in mutations is irrelevant because it can be explained by a single event accounting for all observed mutations so the cost = 1. For example:

A aaa---

B atcttt

C aaattt



In tree (A, (B, C) I can define the subsequence "x" (in red) involving the first three nucleotides and postulate a single mutation event expanding two nucleotides (second and third positions of the alignment) as an autapomorphy on the branch of terminal B so the cost = 1, the same way that the subsequence involving the last three positions of the alignment "y" are explained by a single indel event expanding three nucleotides and having a cost = 1. As I explained above any combination of subsequences is biologically plausible, such as inversions or transpositions of series of nucleotides and loops of ribosomal genes. However, you defend that sets of nucleotides such as those of subsequence "x" should be accounted at the atomistic level (counting every column and transformation independently), while those including indels (e.g., subsequence "y") should be interpreted as single events.

In summary, your approach is inconsistent. For certain sets of data (indels) you apply a *naïve* anti-superfluity principle excluding character independence to calculate similarities, while for other types (nucleotides) you apply a *sophisticated* version of anti-superfluity principle considering character independence. I do not see how your approach is logically more consistent with the principle of explanation of similarities than that of Grant & Kluge with the principle of anti-superfluity. At worst, to me they seem to be equally inconsistent, perhaps the difference is that Grant & Kluge are not thinking about anti-superfluity in the absence of character independence.

Character Independence

One of the core assumptions of phylogenetics is that characters should be independent. We suspect, and sometimes even know, that this assumption is often violated (as with the case of the assumption of hierarchical evolution when species originate by hybridization). Clear examples of non-independent characters are a phenotypic character and the gene that codes for it, several phenotypic characters that evolve as a module, nucleotides that mutate simultaneously such as in ribosomal genes, and a single indel affecting more than one nucleotide but coded as several independent indels. The issue of independence is more complicated than this, but suffices to leave it at this stage to argue my case.

My point is that in most cases it becomes empirically impossible to test the independence of certain sets of characters. DNA sequences are a good example. Let's say we have these two sequences:

ATGCCA

GTACAC

Shall we explain the differences between the first and third position of the alignment as two independent point mutations or as a single event inversion? The same applies to positions 5 and 6 of the alignment. This is just a simple example. It can be much more complicated but I guess it is easy to grasp that the two options are biologically possible and non-testable (at least in this context).

What stops us from explaining everything as a single transformation event is the assumption of character independence. Of course, if we *know* that two or more characters are not independent (*e.g.*, a gene and the phenotype coded by the gene) we should exclude one of the characters. In the absence of evidence, we are better with keeping our assumption; otherwise we open Pandora's box of subjectivity to justify why some types of character sets should be explained by a single event, while others should be considered as several point transformations. For example, consider these sequences:

ATG- - -CCA

GTATTTCAC

What's the scientific reason behind explaining the three indels as a single event expanding three nucleotides but explaining the mutations as single independent events? I see none. So once we let subjectivity in to sweep away character independence, what stop us from asking: Why not to explain all the differences observed between the two sequences as a single event where the polymerase screwed up and all the mutations and indels happen simultaneously?

Explanation and Similarity

Grant & Kluge (2004, 2009), Kluge (2005, 2007), and Kluge & Grant (2006) provide a detailed discussion of why similarity is at best described but not explained, using an object criterion of similarity is logically incompatible with evolutionary theory (classes cannot change/evolve), homoplasy cannot be a hypothesis (ad hoc or not) because it explains nothing, transformations (the events) are the things to be explained (the evidence) and not the objects. I have little to add to their careful discussions, except that in my opinion there is always a gap between the conceptual and the operational but the fact that our operations are doomed to fail in some cases should not stop us from investigating. Operationally we work with and code objects as delimiters of events, but doesn't mean that what we are trying to explain is the similarity (or not) of the objects. A very good discussion on the distinction between the operational and the conceptual is that of Frost & Kluge (1994) *Cladistics*, 10, 259-294 using species concepts and species discovery operations. I see your focus on explaining similarities as an attempt to develop the concepts from the methods, while in my opinion it should be the other way around.